SCIENCE CHINA
Life Sciences

SPECIAL TOPIC
• REVIEW •

February 2013 Vol.56 No.2: 104–109
doi: 10.1007/s11427-013-4441-0

Impact of the next-generation sequencing data depth on various biological result inferences

HOU Rui, YANG ZhenXing, LI MingHui & XIAO HuaSheng*

National Engineering Center for Biochip at Shanghai, Shanghai 201203, China

Received December 29, 2012; accepted January 8, 2013

Next-generation sequencing (NGS) technologies have revolutionized the field of genomics and provided unprecedented opportunities for high-throughput analysis at the levels of genomics, transcriptomics and epigenetics. However, the cost of NGS is still prohibitive for many laboratories. It is imperative to address the trade-off between the sequencing depth and cost. In this review, we will discuss the effects of sequencing depth on the detection of genes, quantification of gene expression and discovering of gene structural variants. This will provide readers information on choosing appropriate sequencing depth that best meet the needs of their particular project.

next-generation sequencing, depth, gene discovery, gene expression, structure variation

Citation: Hou R, Yang Z X, Li M H, et al. Impact of the next-generation sequencing data depth on various biological result inferences. Sci China Life Sci, 2013, 56: 104−109, doi: 10.1007/s11427-013-4441-0

The capillary electrophoresis-based Sanger sequencing method [1] of sequencing by capillary electrophoresis is considered as a 'first-generation' technology, which has been employed in many whole genome sequencing projects and is considered as the 'gold standard' in terms of both read length and sequencing accuracy. In the relatively short time frame since 2005, several next-generation sequencing (NGS) technologies (also known as massively parallel sequencing) have emerged, including Roche 454, Illumina Hiseq 2000, ABI SOLiD, as well as single molecule sequencing technology of Helicos Biosciences and Pacific Biosciences [2], which could make it possible for even single research groups to generate large amounts of sequence data very rapidly and at a substantially lower cost than Sanger method on the ABI 3730xl platform. The significant advances in NGS are revolutionizing the field of genomics and provided unprecedented opportunities for high-throughput analysis at the level of genomics, transcriptomics and epigenetics.

NGS technologies have increased high-throughput capacity and reduced the cost, which makes it possible to use these new platforms for *de novo* sequencing of entire genomes of many species. To date, whole-genome sequencing of many non model species from microbes, plants to animals are available [3–5], facilitated by NGS approaches. For those organisms with a previously published reference genome sequence, NGS can be applied to their whole-genome resequencing [6] for variant discovery, such as single-nucleotide polymorphisms (SNPs), insertions/deletions (InDels), and structural rearrangements. Besides, methods like RRL [7] and RAD [8,9] also allow for discovery of mutations in any non model organisms that are lack of genomic information. Apart from whole genome sequencing, there is much more interest in applying NGS platforms, coupled with DNA target capture technology [10] for focused analysis of specific genomic regions, such as intervals identified through single nucleotide polymorphism (SNP)-based association or linkage studies, candidate genes or the

*Corresponding author (email: huasheng_xiao@shbiochip.com)

whole exome.

In addition to DNA sequencing, the massively parallel sequencing can be applied to sequence RNA, known as RNA-Seq [11]. Microarrays [12] have been recognized as the predominant method for large-scale studies at gene expression levels in past years. In comparison, RNA-Seq is a more powerful tool for comprehensive characterization of whole transcriptome at both gene and exon levels and with an additional ability to identify rare transcripts, new genes, novel splicing junctions and gene fusions. And because RNA-Seq approach does not require the knowledge of the genome sequence as a prerequisite, non model organisms could obtain transcripts information on the sequence level. Furthermore, analysis of small RNAs and other noncoding RNAs with massively parallel sequencing are likely to elucidate the roles of those molecules in gene regulation [13].

Epigenetics is to study heritable gene regulation that does not involve the DNA sequence. Modified NGS protocols also allow for an unbiased assessment of DNA/histone modifications and DNA-protein interactions. The genome-wide single-base resolution of DNA methylation mapping has been performed using bisulfate sequencing [14,15]. Chromatin modifying protein binding can be mapped using chromatin immunoprecipitation sequencing (ChIP-Seq) [16].

The cost of NGS is still prohibitive for many laboratories. NGS platforms allow adding sample specific barcode (sometimes called index) to the library molecules. Then it would be cost-effective if multiple samples are multiplexed and sequenced in a single lane with appropriate sequencing depth. Therefore, one of the most important concerns about a particular NGS experiment is the requirement of sequencing depth. Here, we will discuss the effects of sequencing depth on the detection of genes, quantification of gene expression and discovering of structural variants. This will provide readers information on choosing appropriate sequencing depth that best suit the needs of their particular research.

1 Effect of sequencing depth on gene discovery

RNA-Seq has become a popular method for extracting the transcriptome information. The more the target is sequenced, the more genes are identified. This will presumably result in a more accurate estimation of the expression level. As a consequence, the ability to find genes and detect differential expression is determined by sequencing depth. Low sequencing depth only allows the detection of highly expressed genes. Deeper sequencing is required for those transcripts that are expressed at lower levels. This leads to the question of how many reads should be sufficient to cover the transcriptome of a given sample.

To address this question, we randomly sampled 1–116 million reads from the total reads of three human samples (Figure 1). The number of genes detected was increasing rapidly, and reached the turning point at 20 million reads, with about 45% of ENSEMBL genes were identified. The gene number increased much slowly when the read size was larger than 20 million. The saturation plots were almost identical across the data from sample A and B. In comparison, sample C needed more reads to reach the same percentage of gene detected. The main reason behind the difference was the different RNA preparation method. Sample A and B used poly-A mRNA for sequencing while sample C used the total RNA with rRNA depletion. Non-coding RNA in sample C therefore has a clear influence in the relative proportion of protein coding genes.

Similarly, Huang et al. RNA-Seq analyses for 10 matched pairs of cancer and normal tissues from HBV-related hepatocellular carcinoma (HCC) patients [17]. They randomly sampled 5–65 million reads from the total raw reads from sample A39P and A39C separately. About 15 million raw reads were sufficient to identify more than 50% ENSEMBL genes. This number almost equalled to all detectable expressed genes in a given human tissue. Another research [18] divided their 879 million 50 bp read data from RNA-Seq of cultured human B-cells set into smaller sets and analyzed how the detection of a gene varies with increasing sequencing depth. The total 879-million-read data was assumed to give a comprehensive catalog of transcribed genes. Then the number of genes detected was assessed in fractions of total reads (final data). The number of genes detected reached a plateau at 50 million reads, with about 75% of genes detected. With 100 million reads, 81% of genes were detected. For each additional 100 million reads, 3% more genes were able to be detected on average. Furthermore, expression level of genes also affects the number of genes detected. With 100 million reads, 80% of highly expressed genes (top 25th percentile) compared to 32% of the low expression genes (bottom 25th percentile) were detected.

Additional genes were able to be detected with more reads. However, once the saturation curve excess the turning point, increment of reads results in only a small number of additional genes. To address the trade-off between the

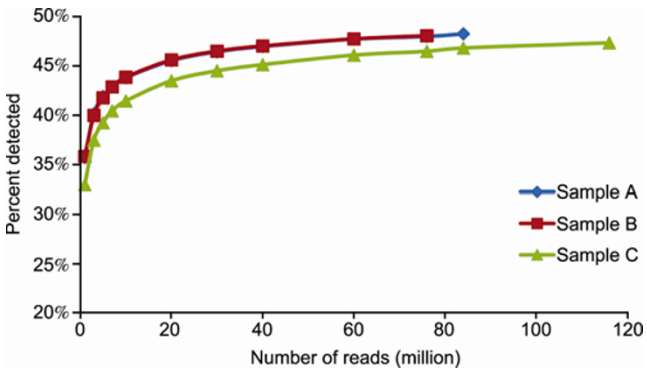

Figure 1 Number of genes detected at different sequencing depths.

depth of RNA-Seq and the coverage of the transcriptome in an organism, Tarazona et al. [19] used new detections rate (NDR), the number of newly detected genes in 1 million additional reads, as a function of the sequencing depth for each of the three published data. Although all three experiments saturation is not entirely reached, each data set has a different sequencing depth; NDRs at the highest depth are substantially different. The NDR value dropped from 232 to 70, and then to 19, with 22, 45 and 200 million reads respectively. In addition, NDR values are broadly similar across data sets for a given number of reads, suggesting that these saturation curves could be indicative for other human data sets generated by Illumina platform.

2 Effect of sequencing depth on quantify gene expression

If only the number of genes was concerned, 15 or 50 million reads could be an answer according to the above discussion. However, for most studies, accurate estimating of expression level is more important. Toung's data [18] gave us a good advice. The expression values assessed by total 879 million reads were set to be the final levels, the sequencing depth necessary to achieve these final levels was analyzed then. At least 400 million reads are needed, when 50% of the genes are required to be within 10% of their final measurements. This is a too large number for most researchers.

If we perform a RNA-Seq experiment using pair-end sequencing (2×100 bp) approach with 4 G per sample (equals to 20 M reads in single-end sequencing like 1×50 bp). With this popular depth for most studies at present, only 1%–2% of genes will have values that are within 10% of their final value. Even with 100 million reads, 6% of genes are within this margin of error. Furthermore, although 100 million reads is sufficient for detection of the majority of genes (72%), the expression levels of these genes deviate from their final value by 41% on average. In other words, if we only consider those highly expressed genes (50% of the total number), and we require these genes to be within 20% of the final measurements, about 170 million reads are needed. Another two studies by Mortazavi et al. [20] and Xu et al. [21] also illuminated that deeper sequencing depth was required for more accurate estimation of low expression genes.

3 Effect of sequencing depth on structural variant discovery

Sequencing depth directly affects the accuracy of expression level estimation, and also affects the structural variant discovery. Alternative splicing is an essential mechanism for increasing transcriptome plasticity and proteome diversity in eukaryotes. About 95% of human genes have alternative splicing, evaluated by a high-throughput sequencing results [22]. As a consequence, the number of splice junctions detected in an experiment close relate to sequencing depth just as the number of genes do. With 100 million reads, 90% of transcripts and 76% of splicing junctions were detected (Figure 1). For each additional 100 million reads, 1% more transcripts and 4% more junctions were able to be detected on average. Furthermore, sufficient sequence coverage is also necessary to estimate the expression of different spliced isoforms and determine their relative abundance. At deficient sequencing depth, low expressed isoforms have large deviation from the final level. What is even worse that the relative abundance between isoforms could be misjudged in some situation [18].

The relative high false positive rate of variants such as SNVs and InDels discovered by RNA-Seq is still a problem [23]. In comparison, whole exome capture sequencing is one of the most frequency applied technique in SNV discovering and subsequently identification of disease-associated alleles. The average depth has increased about three times from this method been applied. We have reviewed 29 publications of whole exome sequencing (Table 1), and found the common depth of exome seqencing was about 30–40× two years ago and about 60–100× recently.

The depth bias in different exon area is the major problem encountered in exome seqencing, as show in Figure 2, about 20% exon region are less than 20× when the average depth is about 50×. The lower the depth is, the higher FDR would exist in SNV and InDel detection. Allele bias is a neglected problem caused by different hybridization efficiency for reference sequence and mutant sequence. The mutant fragment has less opportunity than the right sequence which is perfect match of the probe. It is difficult to calculate the reference-bias for the effect of GC content and hybridization Tm [53].

Allelic balance (AB) was calculated by determining the ratio of reference base calls over the total number of calls at every SNV (AB=0.53–0.55) [54]. We apply ABR (Allele Bias Rate), ABR=ref depth/alt depth, to study whether the bias of InDel is higher than SNV. We also calculate the change of SNV/InDel number and ABR along with depth increment using an exome data (186 M mapped reads) in order to estimate the proper depth sufficient (Figure 3). The result shows that the number of SNV in whole genome keep rising with the depth increment and much higher than SNV number in exon, which saturated at 10 G mapped data, which means that the low depth region outside exon result in a lot of false positive. The InDel number in exon or whole genome saturated at 10 G mapped data, which means that the accuracy of InDel in lower coveraged region is much higher than SNV.

ABR of InDel is higher than SNV because InDel forms loop structure and reduces the stability of the double strands. ABR of SNV and InDel change slightly from 1.2 to 1.3

Table 1 Sequencing depth in publications[a)]

Date	Sequencing platform	Capture platform	Depth/Sample	Reference
Nov 2009	GA II	Agilent 244 K microarrays	40×	[24]
May 2010	SOLiD	Agilent	43×	[25]
Jan 2011	GA II+HiSeq 2000	Agilent 38 M	116×	[26]
Apr 2011	GA IIx 2×76	Agilent	16 G, ~180×	[27]
Jun 2011	GA II 2×78	Agilent	Two lanes (*about 10 G, 100-130×)	[28]
Jul 2011	GAII	Agilent 38 M	11.7–13.4 G	[29]
Aug 2011	GA II 2×75	Agilent 38 M	98×	[30]
Aug 2011	HiSeq 2000 2×50	Agilent	7.3 G (*about 70–900×)	[31]
Aug 2011	GA II 2×101	Agilent	One lane (*about 6–7 G, 60–80×)	[32]
Dec 2011	GA II	Agilent 50 M	~65×	[33]
Dec 2011	GA IIx 2×76	SeqCap EZ	1.4-5.7 G, 27–163×	[34]
May 2012	GA II+HiSeq 2000	Agilent	~80×	[35]
May 2012	GA II 1×75	Agilent's Custom design	32×	[36]
May 2012	HiSeq	Agilent 38 M	118×	[37]
Jun 2012	SOLiD 4	Agilent 50 M	1/4 run	[38]
Jun 2012	HiSeq 2000 2×100	Agilent 50 M	93% target >30×	[39]
Jul 2012	GA II+HiSeq 2000	Agilent 50 M	90×	[40]
Jul 2012	HiSeq 2000 2×101	Agilent 50 M	30–35 G	[41]
Jul 2012	SOLiD 4	Agilent 50 M	43×	[42]
Jul 2012	GA II 2×76	Nimblegen SeqCap EZ	86×	[43]
Jul 2012	HiSeq 2000 2×76	NA	103×	[44]
Jul 2012	GA IIx + HiSeq 2000 2×76	Nimblegen SeqCap EZ	65±14.8×	[45]
Aug 2012	HiSeq 2000	Agilent 38 M	>100× (>11 G)	[46]
Aug 2012	GA II 1×75+SOLiD	NimbleGen+Agilent	28×, 88×	[47]
Sept 2012	GA II 2×95	agilent 38 M	190×	[48]
Sept 2012	HiSeq 2000 2×75	Agilent 38 M+50 M	80×, 162×	[49]
Sept 2012	HiSeq 2000	Agilent 38 M	56×	[50]
Sept 2012	GA II 2×75	Agilent 38 M	6.9 G (*about 60–80×)	[51]
Oct 2012	HiSeq 2000 2×100	Agilent	63×	[52]

a) *, Estimates.

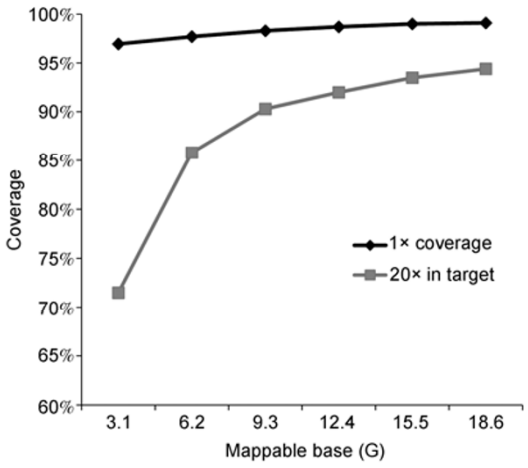

Figure 2 Coverage of 1× and 20× regions at different sequencing depths.

when depth increasing, which means that the bias is common but slight. The increment of InDel ABR may be caused by new InDels detected with lagre ABR when depth increasing. In summary, 10 G of mapped data (average sequencing depth about 100×) will be appropriate for exome seqencing, which matches the current trend.

4 Conclusion

In this study, we have discussed the effects of sequencing depth on detection of genes, quantification of gene expression and discovering of structural variants. The results indicated that although RNA-Seq effectively enhances our view on the diversity of the transcriptome, the quantification of true expression at a low depth might not be so easy to achieve. In practice, 15–50 million reads allow to detect the majority of genes in human tissue. However, at least 170 million reads are needed, when 50% of the genes were required to be within 20% of their final measurements. For the detection of splicing junctions, similar results were reached. For variants such as SNVs and InDels, exome se-

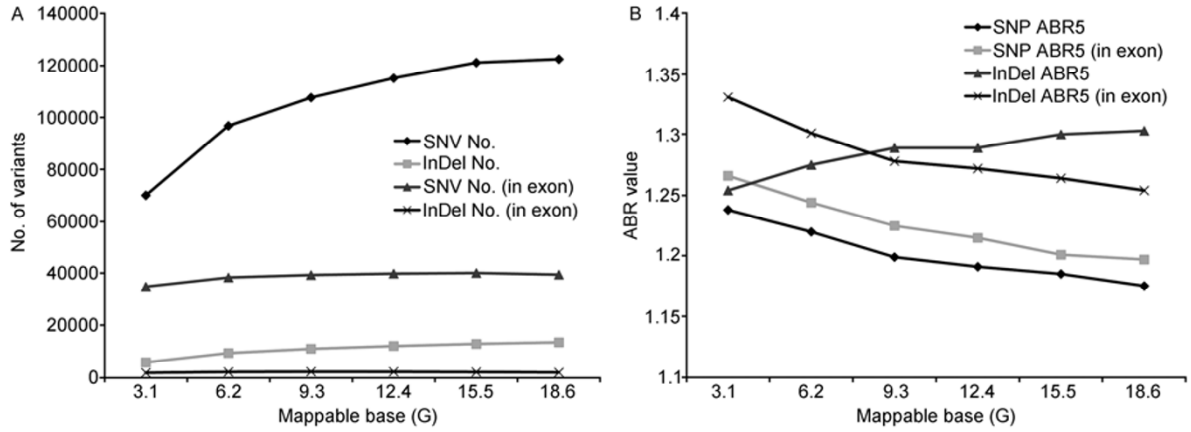

Figure 3 Number of varients (A) and ABR value (B) at different sequencing depths.

quencing is a powerful approach, and 10 G of mapped data (about 100×) will be appropriate for most studies, which matches the current trend.

1 Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, 74: 5463–5467
2 Metzker M L. Sequencing technologies—the next generation. Nat Rev Genet, 2010, 11: 31–46
3 Huang S, Li R, Zhang Z, et al. The genome of the cucumber, *Cucumis sativus* L. Nat Genet, 2009, 41: 1275–1281
4 Li R, Fan W, Tian G, et al. The sequence and de novo assembly of the giant panda genome. Nature, 2010, 463: 311–317
5 Hernandez D, Francois P, Farinelli L, et al. *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res, 2008, 18: 802–809
6 Wheeler D A, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature, 2008, 452: 872–876
7 van Tassell C P, Smith T P, Matukumalli L K, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods, 2008, 5: 247–252
8 Baird N A, Etter P D, Atwood T S, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE, 2008, 3: e3376
9 Wang S, Meyer E, McKay J K, et al. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat Methods, 2012, 9: 808–810
10 Ng S B, Turner E H, Robertson P D, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 2009, 461: 272–276
11 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
12 Fodor S P, Rava R P, Huang X C, et al. Multiplexed biochemical assays with biological chips. Nature, 1993, 364: 555–556
13 Morin R D, O'Connor M D, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res, 2008, 18: 610–621
14 Lister R, O'Malley R C, Tonti-Filippini J, et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell, 2008, 133: 523–536
15 Ordway J M, Budiman M A, Korshunova Y, et al. Identification of novel high-frequency DNA methylation changes in breast cancer. PLoS ONE, 2007, 2: e1314
16 Johnson D S, Mortazavi A, Myers R M, et al. Genome-wide mapping of in vivo protein-DNA interactions. Science, 2007, 316: 1497–1502
17 Huang Q, Lin B, Liu H, et al. RNA-Seq analyses generate comprehensive transcriptomic landscape and reveal complex transcript patterns in hepatocellular carcinoma. PLoS ONE, 2011, 6: e26168
18 Toung J M, Morley M, Li M, et al. RNA-sequence analysis of human B-cells. Genome Res, 2011, 21: 991–998
19 Tarazona S, Garcia-Alcalde F, Dopazo J, et al. Differential expression in RNA-seq: a matter of depth. Genome Res, 2011, 21: 2213–2223
20 Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
21 Xu W, Seok J, Mindrinos M N, et al. Human transcriptome array for high-throughput clinical studies. Proc Natl Acad Sci USA, 2011, 108: 3707–3712
22 Pan Q, Shai O, Lee L J, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, 2008, 40: 1413–1415
23 Cirulli E T, Singh A, Shianna K V, et al. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. Genome Biol, 2010, 11: R57
24 Ng S B, Buckingham K J, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet, 2010, 42: 30–35
25 Hoischen A, van Bon B W, Gilissen C, et al. *De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet, 2010, 42: 483–485
26 Wang K, Kan J, Yuen S T, et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat Genet, 2011, 43: 1219–1223
27 Wei X, Walia V, Lin J C, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet, 2011, 43: 442–446
28 Comino-Mendez I, Gracia-Aznarez F J, Schiavi F, et al. Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. Nat Genet, 2011, 43: 663–667
29 Albers C A, Cvejic A, Favier R, et al. Exome sequencing identifies NBEAL2 as the causative gene for gray platelet syndrome. Nat Genet, 2011, 43: 735–737
30 Li M, Zhao H, Zhang X, et al. Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat Genet, 2011, 43: 828–829
31 Xu B, Roos J L, Dexheimer P, et al. Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. Nat Genet, 2011, 43: 864–868
32 Sloan J L, Johnston J J, Manoli I, et al. Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria. Nat Genet, 2011, 43: 883–886
33 Quesada V, Conde L, Villamor N, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic

lymphocytic leukemia. Nat Genet, 2011, 44: 47–52
34 Nikolaev S I, Rimoldi D, Iseli C, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet, 2011, 44: 133–139
35 Ong C K, Subimerb C, Pairojkul C, et al. Exome sequencing of liver fluke-associated cholangiocarcinoma. Nat Genet, 2012, 44: 690–693
36 Arboleda V A, Lee H, Parnaik R, et al. Mutations in the PCNA-binding domain of CDKN1C cause IMAGe syndrome. Nat Genet, 2012, 44: 788–792
37 Barbieri C E, Baca S C, Lawrence M S, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet, 44: 685–689
38 Wortmann S B, Vaz F M, Gardeitchik T, et al. Mutations in the phospholipid remodeling gene SERAC1 impair mitochondrial function and intracellular cholesterol trafficking and cause dystonia and deafness. Nat Genet, 2012, 44: 797–802
39 Lee J H, Huynh M, Silhavy J L, et al. *De novo* somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat Genet, 2012, 44: 941–945
40 Heinzen E L, Swoboda K J, Hitomi Y, et al. *De novo* mutations in ATP1A3 cause alternating hemiplegia of childhood. Nat Genet, 2012, 44: 1030–1034
41 Falk M J, Zhang Q, Nakamaru-Ogiso E, et al. NMNAT1 mutations cause Leber congenital amaurosis. Nat Genet, 2012, 44: 1040–1045
42 Harakalova M, van Harssel J J, Terhal P A, et al. Dominant missense mutations in ABCC9 cause Cantu syndrome. Nat Genet, 2012, 44: 793–796
43 Emond M J, Louie T, Emerson J, et al. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. Nat Genet, 2012, 44: 886–889
44 Hodis E, Watson I R, Kryukov G V, et al. A landscape of driver mutations in melanoma. Cell, 2012, 150: 251–263
45 Krauthammer M, Kong Y, Ha B H, et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat Genet, 2012, 44: 1006–1014
46 Rademakers R, Baker M, Nicholson A M, et al. Mutations in the colony stimulating factor 1 receptor (CSF1R) gene cause hereditary diffuse leukoencephalopathy with spheroids. Nat Genet, 2012, 44: 200–205
47 Huang J, Deng Q, Wang Q, et al. Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. Nat Genet, 2012, 44: 1117–1121
48 Peifer M, Fernandez-Cuesta L, Sos M L, et al. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet, 2012, 44: 1104–1110
49 Rudin C M, Durinck S, Stawiski E W, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet, 2012, 44: 1111–1116
50 Rice G I, Kasher P R, Forte G M, et al. Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. Nat Genet, 2012, 44: 1243–1248
51 Doyle A J, Doyle J J, Bessling S L, et al. Mutations in the TGF-beta repressor SKI cause Shprintzen-Goldberg syndrome with aortic aneurysm. Nat Genet, 2012, 44: 1249–1254
52 Barcia G, Fleming M R, Deligniere A, et al. *De novo* gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. Nat Genet, 2012, 44: 1255–1259
53 Bainbridge M N, Wang M, Burgess D L, et al. Whole exome capture in solution with 3 Gbp of data. Genome Biol, 2010, 11: R62
54 Clark M J, Chen R, Lam H Y, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol, 2011, 29: 908–914